Detecting hallucinations in large language models using semantic entropy

Year: 2024 | *Citations: 733* | *Authors: Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn*

Abstract

Large language model (LLM) systems, such as ChatGPT1 or Gemini2, can show impressive reasoning and question-answering capabilities but often ‘hallucinate’ false outputs and unsubstantiated answers3,4. Answering unreliably or without the necessary information prevents adoption in diverse fields, with problems including fabrication of legal precedents5 or untrue facts in news articles6 and even posing a risk to human life in medical domains such as radiology7. Encouraging truthfulness through supervision or reinforcement has been only partially successful8. Researchers need a general method for detecting hallucinations in LLMs that works even with new and unseen questions to which humans might not know the answer. Here we develop new methods grounded in statistics, proposing entropy-based uncertainty estimators for LLMs to detect a subset of hallucinations—confabulations—which are arbitrary and incorrect generations. Our method addresses the fact that one idea can be expressed in many ways by computing uncertainty at the level of meaning rather than specific sequences of words. Our method works across datasets and tasks without a priori knowledge of the task, requires no task-specific data and robustly generalizes to new tasks not seen before. By detecting when a prompt is likely to produce a confabulation, our method helps users understand when they must take extra care with LLMs and opens up new possibilities for using LLMs that are otherwise prevented by their unreliability. Hallucinations (confabulations) in large language model systems can be tackled by measuring uncertainty about the meanings of generated responses rather than the text itself to improve question-answering accuracy.